

# PROBABILISTIC CROSS-IDENTIFICATION OF ASTRONOMICAL SOURCES

TAMÁS BUDAVÁRI AND ALEXANDER S. SZALAY

Dept. of Physics and Astronomy, The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218 and  
 Max-Planck-Institute für Astrophysik, Karl-Schwarzschild-Strasse 1, 85748 Garching, Germany

*Draft version February 11, 2008*

## ABSTRACT

We present a general probabilistic formalism for cross-identifying astronomical point sources in multiple observations. Our Bayesian approach, symmetric in all observations, is the foundation of a unified framework for object matching, where not only spatial information, but physical properties, such as colors, redshift and luminosity, can also be considered in a natural way. We provide a practical recipe to implement an efficient recursive algorithm to evaluate the Bayes factor over a set of catalogs with known circular errors in positions. This new methodology is crucial for studies leveraging the synergy of today's multi-wavelength observations and to enter the time-domain science of the upcoming survey telescopes.

*Subject headings:* astrometry — catalogs — galaxies: statistics — methods: statistical

## 1. MOTIVATION

Observational astronomy has changed dramatically over the last decade. With the introduction of large-format, high-resolution detectors at all wavelengths of the electromagnetic spectrum, astronomers now face an avalanche of data pouring from the instruments of dedicated telescopes. While most imaging surveys today obtain multicolor information, no one telescope can cover the entire spectrum because the physics of the detectors is very different at different frequencies. To fully utilize the available observations, e.g., to boost the chances of discovering new kinds of sources, and understanding the underlying physical relations of object properties in a statistical way, one needs to merge the datasets of various telescopes by federating the archives. The Virtual Observatory initiative spearheaded by the International Virtual Observatory Alliance<sup>1</sup> (IVOA) is pursuing automated data exchange protocols with catalog cross-identification, and the US National Virtual Observatory<sup>2</sup> (NVO) is building tools, e.g., Open SkyQuery (Budavári et al. 2004) to facilitate a standard unified framework. The key step in the process is the cross-identification of the sources in multiple catalogs to link observations of one telescope to other's. Previous attempts to alleviate the problem utilized likelihood analysis (Sutherland & Saunders 1992) and machine learning (ML) techniques (Rohde et al. 2006) that addressed specific issues of the matching problem of two catalogs. Mann et al. (1997) successfully applied the former likelihood ratio method to associate sources in the Infrared Space Observatory and the Hubble Deep Field catalogs, and the ML techniques were used to study the SuperCOSMOS observations and HI Parkes All Sky Survey.

Today astronomers typically join two catalogs by setting some threshold on the angular separation of sources that is motivated by the astrometric accuracies of the datasets involved. When more than two catalogs are to be crossmatched, astronomers often hatch a chaining rule based on the implicit prior knowledge about the sources. For example, one might decide to match all lower-accuracy datasets to the best one, or to go from wavelength to wavelength, hoping that the sources do not change significantly over a shorter wavelength range.

The problem with these traditional ways is not that they are based on implicit assumptions and intuitions but that they are not symmetric. While the pairwise matches might be acceptable, there is no guarantee, or any measure of quality, that the elected final matches are plausible or if the list is complete. After all picking a different order of pairwise matching would yield a different catalog.

We need algorithms that are symmetric in the catalogs and provide a reliable measure of quality that one can use to exclude or downweight unlikely combinations of sources. We need a unified framework, where on top of the spatial information, other measurements can also be incorporated along with explicit models and physical priors. In Section 2 we discuss the Bayesian approach to address these issues, and in Section 3 the spherical normal distribution is studied. In Sections 4 we demonstrate how to include other observational evidence such as from multicolor photometric measurements. Section 5 focuses on the effects of a limited field of view on the observational evidence, the prior and posterior probabilities. In Section 6 an efficient implementation of the framework is described in the detail, and Section 7 concludes our study.

Throughout the paper we follow the usual convention of using the lower-case  $p$  symbol for representing probability density functions and the capital  $P$  symbol for probabilities.

## 2. OBSERVATIONAL EVIDENCE

Often Bayesian analysis is referred to as the calculus of belief, however, it should rather be thought of as the calculus of observational evidence. When presented with a series of observed positions, one would like to know whether they are truly from the same source. If the coordinates are scattered all over the celestial sphere, it seems very unlikely that they are measurements of the same astronomical object, but when the coordinates are only a tiny fraction of an arcsecond apart, we “know” that we found a good match. How good is that match? Or what is the evidence that it is a match?

### 2.1. Modelling the Astrometry

First let us examine what astrometric precision means. In the process of calibrating the positions in a catalog of extracted sources, one can characterize the properties of the observations by comparing the positions to astrometric stan-

<sup>1</sup> <http://www.ivoa.net>

<sup>2</sup> <http://us-vo.org>

dards, and even correct for systematic offsets. Yet, there remains a random scatter around the true positions. This uncertainty is often modelled as a normal distribution, and catalogs would quote a single  $\sigma$ -value for their accuracy, e.g.,  $\sigma = 0.1$  arcseconds. In general, our understanding of the astrometry is described by a probability density function (PDF) that may even vary on the sky. We parameterize our model  $M$  that the object is on the celestial sphere using a three-dimensional normal vector  $\vec{m}$ , and write  $p(\vec{x}|\vec{m}, M)$  for the probability density that an object at its true location  $\vec{m}$  is observed at a position  $\vec{x}$ . As any PDF, this function is normalized,

$$\int p(\vec{x}|\vec{m}, M) d^3x = 1 \quad (1)$$

Now we take a single source observed at  $\vec{x}_1$  and apply Bayes' theorem to find the posterior density of the true location  $\vec{m}$  given the obtained data,

$$p(\vec{m}|\vec{x}_1, M) = \frac{p(\vec{x}_1|\vec{m}, M)p(\vec{m}|M)}{p(\vec{x}_1|M)} \quad (2)$$

where the trivial prior  $p(\vec{m}|M)$  of  $\vec{m}$  being on the celestial sphere is expressed with Dirac's  $\delta$ -symbol,

$$p(\vec{m}|M) = \frac{1}{4\pi} \delta(|\vec{m}| - 1) \quad (3)$$

and the normalizing constant guarantees the law of total probability,

$$p(\vec{x}_1|M) = \int p(\vec{m}|M) p(\vec{x}_1|\vec{m}, M) d^3m \quad (4)$$

Another interesting direct application is the calculation of the chance that we find a visible object at position  $\vec{m}$  in a given footprint. If the angular window function is  $\Omega$ , this probability is simply

$$P(\Omega|\vec{m}, M) = \int_{\Omega} p(\vec{x}|\vec{m}, M) d^3x \quad (5)$$

which one can use to infer the PDF on the true position by applying Bayes' rule

$$p(\vec{m}|M_{\Omega}) \equiv p(\vec{m}|\Omega, M) = \frac{p(\vec{m}|M)P(\Omega|\vec{m}, M)}{\int p(\vec{m}|M)P(\Omega|\vec{m}, M) d^3m} \quad (6)$$

This is our best understanding of where an object might be on the sky (prior to measuring its actual position) that is seen in the specified  $\Omega$  footprint assuming astrometric precision  $p(\vec{x}|\vec{m}, M)$  derived from the calibration.

### 2.2. The Bayes Factor

With multiple observations through various instruments of possibly different astrometric accuracies, we now turn to compute the evidence that all observations are from the same source. We introduce the Bayes factor to test this hypothesis  $H$  against the case when separate sources are possible,  $K$ . After the observation are obtained,  $D = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}$  locations on the sky, we compute the ratio of the posterior and prior probabilities of each hypothesis. The Bayes factor is defined as the ratio of these odds,

$$B(H, K|D) = \left( \frac{P(H|D)}{P(H)} \right) / \left( \frac{P(K|D)}{P(K)} \right) \quad (7)$$

which, after applying Bayes' theorem, becomes

$$B(H, K|D) = \frac{p(D|H)}{p(D|K)} \quad (8)$$

for continuous observables. The actual calculation is done by parameterizing the two models  $H$  and  $K$ , and integrating the likelihood functions for the entire configuration space. Our hypothesis  $H$  says that the positions are from a single source, thus can be parameterized by a single common location  $\vec{m}$ . Due to the independence of the measurements in  $D$ , the joint PDF is just the product of the astrometric precisions  $p_1, p_2, \dots, p_n$ , and the integral simplifies to

$$p(D|H) = \int p(\vec{m}|H) \prod_{i=1}^n p_i(\vec{x}_i|\vec{m}, H) d^3m \quad (9)$$

On the other hand, the alternative hypothesis  $K$  is parameterized by separate  $\{\vec{m}_i\}$  positions, and the integral factorizes into the product of the independent components

$$p(D|K) = \prod_{i=1}^n \left\{ \int p(\vec{m}_i|K) p_i(\vec{x}_i|\vec{m}_i, K) d^3m_i \right\} \quad (10)$$

When the Bayes factor is large, the observations support the hypothesis that the association is a match, if it is in the order of unity, the evidence is not convincing, and finally if the ratio is less than one, the data prefers the alternative hypothesis.

### 3. THE NORMAL DISTRIBUTION

Normal distributions emerge often in nature, where a number of effects play roles in shaping up the probability density, cf. the Central Limit theorem. Although many of the usual arguments do not hold over closed topological manifolds, e.g., the Central Limit theorem leads to isotropic distribution on the circle (Lévy 1939), it is possible to introduce an analogue to the normal distribution function on the sphere (Fisher 1953; Breitenberger 1963). The spherical normal distribution is often elected to characterize the precision of astronomy observations, hence it is of great importance to understand its properties, and to apply the Bayesian framework described in the previous section.

The spherical normal distribution in its normalized form using the previous 3-D vector notation is written as

$$N(\vec{x}|\vec{m}, w) = \frac{w \delta(|\vec{x}| - 1)}{4\pi \sinh w} \exp(w \vec{m} \cdot \vec{x}) \quad (11)$$

where the weight  $w$  is typically very large. When this is the case, the weight is related to the more intuitive precision parameter  $\sigma$  by the equation

$$w = 1/\sigma^2 \quad (12)$$

For example, when  $\sigma$  is in the order of an arcsecond, the weight takes values of  $\sim 10^{10}$ . Having observed a set of positions independently with corresponding weights, we can compute the Bayes factor for the two hypotheses  $H$  and  $K$  introduced earlier. Because the function  $N(\vec{x}|\vec{m}, w)p(\vec{m}|M)$  is symmetric in  $\vec{x}$  and  $\vec{m}$  for the trivial prior, and the PDFs are normalized, the Bayes factor is computed analytically, and becomes

$$B(H, K|D) = \frac{\sinh w}{w} \prod_{i=1}^n \frac{w_i}{\sinh w_i} \quad (13)$$

with

$$w = \left| \sum_{i=1}^n w_i \vec{x}_i \right| \quad (14)$$

where we exploit the fact that the product of normal distributions has the same functional form. A detailed derivation is given in Appendix A.

In case of only two observations, this weight depends on the astrometric precisions and the angle  $\psi$  between the positions

$$w = \sqrt{w_1^2 + w_2^2 + 2w_1w_2 \cos \psi} \quad (15)$$

For the typical large weights and small angular separations between the measurements, we get

$$B = \frac{2}{\sigma_1^2 + \sigma_2^2} \exp \left\{ -\frac{\psi^2}{2(\sigma_1^2 + \sigma_2^2)} \right\} \quad (16)$$

In Figure 1 the 10-based logarithm of the Bayes factor, also known as the weight of evidence, is shown as a function of angular separation for the three cases of matching two catalogs of  $\sigma_1 = 0.1''$  and  $\sigma_2 = 0.5''$  to each other and to themselves. This is the problem of matching the Sloan Digital Sky Survey (SDSS; York et al. 2000; Pier et al. 2003) and the Galaxy Evolution Explorer (GALEX; Martin et al. 2005; Morrissey et al. 2007) science archives.

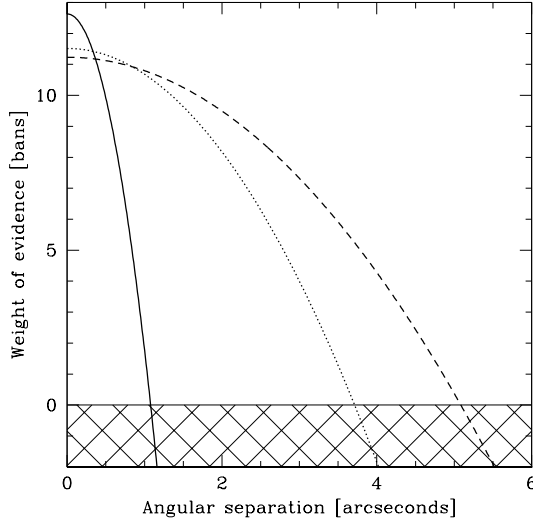


FIG. 1.— The weight of evidence as a function of angular separation for the three cases of matching two catalogs of  $\sigma_1 = 0.1''$  and  $\sigma_2 = 0.5''$  to each other and to themselves. For example, matching SDSS and GALEX sources: SDSS–SDSS (solid), SDSS–GALEX (dotted) and GALEX–GALEX (dashed).

Matching three catalogs also makes an interesting case study for the various potential configurations of three positions. The Bayes factor for this case, in the same limit as previously, takes the form of

$$B = \frac{4 \exp \left\{ -\frac{\sigma_1^2 \psi_{12}^2 + \sigma_1^2 \psi_{23}^2 + \sigma_2^2 \psi_{31}^2}{2(\sigma_1^2 \sigma_2^2 + \sigma_2^2 \sigma_3^2 + \sigma_3^2 \sigma_1^2)} \right\}}{\sigma_1^2 \sigma_2^2 + \sigma_2^2 \sigma_3^2 + \sigma_3^2 \sigma_1^2} \quad (17)$$

In Table 1 the weight of evidence is shown for various configurations from the matching of three similar catalogs with equal astrometric accuracies,  $\sigma_1 = \sigma_2 = \sigma_3 = 0.1''$  (separations listed in  $\sigma$  units.) The astrometric precision was chosen to match the nominal SDSS limitations.

In general, the Bayes factor for the typical large weights and small angular separations takes the form of

$$B = 2^{n-1} \frac{\prod w_i}{\sum w_i} \exp \left\{ -\frac{\sum_{i < j} w_i w_j \psi_{ij}^2}{2 \sum w_i} \right\} \quad (18)$$

where all summations and products run on the members of the tuple from the  $n$  number of catalogs; see Appendix B for the details of the calculation.

TABLE 1  
WEIGHT OF EVIDENCE FOR THREE SURVEYS AS A FUNCTION OF THE ANGULAR SEPARATIONS IN  $\sigma = \sigma_1 = \sigma_2 = \sigma_3 = 0.1''$  UNITS

$\psi_{12}$	$\psi_{23}$	$\psi_{31}$	$W$	$\psi_{12}$	$\psi_{23}$	$\psi_{31}$	$W$
0	0	0	25.38	0	0	0	25.38
1	1	1	25.17	0	1	1	25.24
2	2	2	24.51	0	2	2	24.80
3	3	3	23.43	0	3	3	24.08
4	4	4	21.91	0	4	4	23.07
5	5	5	19.95	0	5	5	21.76
6	6	6	17.57	0	6	6	20.17
7	7	7	14.74	0	7	7	18.29
8	8	8	11.49	0	8	8	16.12
9	9	9	7.79	0	9	9	13.66
10	10	10	3.67	0	10	10	10.91
11	11	11	-0.89	0	11	11	7.87
12	12	12	-5.89	0	12	12	4.54
13	13	13	-11.32	0	13	13	0.92
14	14	14	-17.18	0	14	14	-2.99

In scenarios where individual errors are different or even anisotropic, one can generalize our expression in a fairly straightforward manner in the above approximation. Instead of the scalar weight, one can use the inverse of the covariance matrix, however, the elegant simplicity of the expressions is sacrificed.

#### 4. FOLDING IN THE PHYSICS

Naturally the formalism introduced in Section 2 is not specific to astrometric observations. In fact, it is rather straightforward to fold other measured quantities into the calculations. This is especially important when dealing with multiple matches. Picking the “correct” combination of sources from various spatially similar configurations is a degenerate problem that requires extra information to resolve. The use of photometric information is a natural choice for its wide availability, however, its application requires further assumptions on the spectral energy distributions (SEDs). Often models exist to help out with the solution, but extra caution is needed to avoid any undesirable effect. For example, when the goal is to discover new types of objects with unknown SEDs, one should not apply known SEDs as priors but rather look for combinations that are likely matches based on spatial detections but excluded by SED modelling.

As a demonstration of these ideas, let us apply the introduced Bayesian framework to photometric measurements in various passbands. The ingredients include the following further explicit models:

- 1 Model  $S$  for the spectrum energy distributions, e.g., by Bruzual & Charlot (2003), described by a set of parameters,  $\vec{\eta}$ :  $s(\lambda|\vec{\eta}, S)$  along with the corresponding  $p(\vec{\eta}|S)$  priors;
- 2 Model  $R$  for the transmission of the passbands to calculate simulated fluxes  $\vec{\gamma}(\vec{\eta}|S, R)$  by integrating the SEDs  $s(\lambda|\vec{\eta}, S)$  with the appropriate response functions; and
- 3 Model  $C$  for the uncertainty of the catalog from the photometric calibration,  $p(\vec{g}|\vec{\gamma}, C)$ , where  $\vec{g}$  is the observed flux set and  $\vec{\gamma}$  is the true.

These separate models can be folded into a single model  $M$ , for simplicity, so one can write  $p(\vec{g}|\vec{\eta}, M)$  for the probability density of measuring  $\vec{g}$  fluxes for an object with  $\vec{\eta}$  physical

properties of  $S$  seen through the filters in  $R$  with the  $C$  photometric accuracy. The Bayes factor for the photometry in the face of the observed fluxes  $D' = \{\vec{g}_1, \vec{g}_2, \dots, \vec{g}_n\}$ , similarly to the astrometric formulas, is given by the ratio

$$B(H, K|D') = \frac{\int p(\vec{\eta}|H) \prod_{i=1}^n p_i(\vec{g}_i|\vec{\eta}, H) d^r \eta}{\prod_{i=1}^n \left\{ \int p(\vec{\eta}_i|K) p_i(\vec{g}_i|\vec{\eta}_i, K) d^r \eta_i \right\}} \quad (19)$$

In the simplest case,  $S$  is parameterized by a discrete spectral type  $T$ , the redshift  $z$  and an overall scaling factor for the brightness,  $\alpha$ :

$$\vec{\gamma} = \alpha \vec{f}(T, z) \quad (20)$$

where  $\vec{f}$  is a vector of the simulated photometry in the various passbands. Photometric uncertainties are often assumed to be Gaussian with a diagonal covariance matrix of elements  $\sigma_l^2$ , where  $l$  runs on the  $L$  number of passbands. After substitution, we arrive at the familiar formula of

$$p(\vec{g}|\vec{\eta}, M) = \frac{1}{\mathcal{N}} \exp \left\{ - \sum_{l=1}^L \frac{[g_l - \alpha f_l(T, z)]^2}{2\sigma_l^2} \right\} \quad (21)$$

where constant  $\mathcal{N}$  is the usual normalization factor of the multivariate normal distribution, which in our special case is just  $\mathcal{N} = (2\pi)^{L/2} \sigma_1 \sigma_2 \dots \sigma_L$ . Integrating these models to get the Bayes factor is a very similar problem to template fitting photometric redshift estimation. In fact, the two procedures can be done in a self-consistent way within the same application. Naturally, spectroscopic redshift measurements can be directly incorporated in this analysis, when available, but other data can also enter in a straightforward manner.

The Bayesian analysis is inherently recursive. As soon as we obtain new measurements, and compute the posterior probability, that becomes the prior for subsequent studies. This is an extremely powerful property, and simplifies the computations enormously. A consequence of this is that the combined Bayes factor of the astrometric and photometric measurements is simply the product of the two,

$$B = B_{\text{pos}} \cdot B_{\text{phot}} \quad (22)$$

as also seen from the Bayes factor's definition. This means that one can just do the spatial join first, and consider additional measurements and physical priors in subsequent steps, if needed.

## 5. FROM PRIORS TO POSTERIOR

The Bayes factor naturally relates the prior and posterior probabilities. When  $K$  is the complementary hypothesis of  $H$ , the posterior probability is

$$P(H|D) = \left[ 1 + \frac{1 - P(H)}{BP(H)} \right]^{-1} \quad (23)$$

which, in the limit of vanishing priors, becomes

$$P(H|D) = \frac{BP(H)}{1 + BP(H)} \quad (24)$$

To make a definitive decision on whether a set of detections should be considered a match, one would like to set a limit on the posterior probability and derive the Bayes factor threshold from that, however, this can only be done with an initial estimate of the prior.

### 5.1. The Prior and the Selection Function

The prior probability depends on the angular and radial selection functions of the observations. If the visible universe contains  $N$  objects, and we select two of them at random, the probability of picking the same object is  $1/N$ . When selecting  $n$  objects, the probability is  $1/N^{n-1}$ . A limited field of view shrinks the observable volume, hence decreases the number of objects, and increases the prior probability. When the angular selection functions of the catalogs overlap only partially then one can just consider the intersection of the sky coverage and the smaller number of sources within.

The various radial selection functions also have a significant role, and make the situation more complicated. In order to consider their effect, one has to estimate the overlap of the selections in the input catalogs. Every catalog has observational constraints, other than the field of view, like flux limits, that set the radial selection function. The superset of these constraints defines the restrictions on the *overlap catalog*. Let  $N_*$  denote the number of objects in that catalog. In this general case, the prior probability takes the form of

$$P(H) = N_* / \prod_i^n N_i \quad (25)$$

When the limitations are identical, all catalogs have equal number of objects,  $N_* = N_1 = \dots = N_n$ , and we get back the same formula of  $P(H) = 1/N^{n-1}$  as before, but when, for example, one catalog consists of only low-redshift galaxies (e.g.,  $z < 0.2$ ), and the other has high-redshift quasars (e.g.,  $z > 3$ ), there is no overlap between the two radial selection functions, hence  $N_* = 0$ , which means  $P(H) = 0$ . One can get vanishing priors even if the redshift histograms overlap significantly, e.g., two catalogs of red (e.g.,  $u - g < 2$ ) and blue galaxies ( $u - g > 2$ ). In general, all these complex selection effects are captured in a single scalar quantity,  $N_*$ , which is estimated based on prior physical knowledge, e.g., by using template SEDs and the known characteristics of the input catalogs (e.g., the luminosity functions), or alternatively, when no prior information is available, one can invoke self-consistency arguments to derive it; see later. We now rewrite the prior with the surface densities,  $\nu = N/\Omega$ , or the scaled number of objects for the entire sky,  $\rho = 4\pi\nu$ , as

$$P(H) = \frac{\nu_*}{\prod \nu_i} \Omega^{1-n} = \frac{\rho_*}{\prod \rho_i} \left( \frac{\Omega}{4\pi} \right)^{1-n} \quad (26)$$

This formula also provides a straightforward way to include a model for varying surface density on the sky, e.g., for stars, where  $\nu = \nu(\vec{x})$ . In this case, a constant limiting posterior probability yields a varying threshold on the Bayes factor as a function of the position on the celestial sphere.

### 5.2. The Bayes Factor and the Window Function

The field of view not only changes the prior probabilities but also modifies the Bayes factor. When the window function is known, one can refine the prior probability density that enters the integral of the numerator and denominator of the Bayes factor. This is done by adopting eq. 6 as the prior. In first order, for the typical catalogs with large weights (high accuracy) and large contiguous observation areas, this new prior is uniform over the window function, neglecting the fuzzy boundary, except scaled by the area coverage

$$p(\vec{m}|M_\Omega) = \frac{\Omega(\vec{m})}{\Omega} \delta(|\vec{m}| - 1) \quad (27)$$

where again  $\Omega(\vec{m})$  is the window function that takes the value 1 when  $\vec{m}$  is inside and 0 otherwise, and  $\Omega$  is its area. The Bayes factor inside the footprint will be essentially same as before in eq. 13 but also scaled with these fractional coverage:

$$B = \left( \frac{\Omega}{4\pi} \right)^{n-1} \frac{\sinh w}{w} \prod_{i=1}^n \frac{w_i}{\sinh w_i} \quad (28)$$

The edge effect modifies this only for a tiny fraction of the objects at the boundary of the observations. The proper integral is, of course, much more expensive than this analytical formula, but can be evaluated or re-evaluated, e.g., by an MCMC algorithm.

For the typical small priors, the posterior depends only on the product of the Bayes factor and the prior; see eq. 24. This means that the footprint effect cancels out in the posterior probability; cf. eqs. 26 and 28. Hence it is still sensible to just simply use the all-sky formula in eq. 13 and 18 as long as the prior is written accordingly, i.e.,  $\rho_*/\rho_1\rho_2\cdots\rho_n$ . From the data providers' point of view, who often do not know the field of view, e.g., the legacy catalogs in VizieR (Ochsenbein et al. 2000) or the NASA/IPAC Extragalactic Database (NED; Madore et al. 1992), the best quantity to publish along with the matched tuples is also the analytic all-sky Bayes factor, so researchers can incorporate their own prior knowledge, and set the thresholds on the posterior accordingly that are often specific to the science application.

### 5.3. Self-Consistent Estimation

In principle, the cross-identification process is now complete, one just has to formulate the prior, possibly varying on the sky, and set a threshold on the posterior probabilities to select the matches. For the ignorant without a priori knowledge, these are not completely independent choices, and, at least in the limit when all observables are being considered in the Bayes factors, could be derived from requirements of a self-consistent field theory. When prior knowledge is available and dictates a preference, one could and probably should still check for the consistency outlined here to understand the discrepancies, if any.

The formula for the prior in eq. 25 is in fact equivalent to stating that  $P(H)$  is constant and

$$\sum P(H) = N_* \quad (29)$$

where the summation runs over the direct product of all sets of sources in the  $n$  catalogs, i.e., all possible combinations of detections with  $N_1 N_2 \cdots N_n$  contributions. The self-consistency argument requires that

$$\sum P(H|D) = N_* \quad (30)$$

which is an equation for  $N_*$  that can be solved by, e.g., some iterative approximation method starting from an initial value of  $N_* = \min \{N_i\}$ . Initial experiments support our expectations that these procedures indeed converge very rapidly, only in a few iterations, and are insensitive to the matching limit once the Bayes factor is less than unity. For varying unknown priors one can use some sky tessellation schemes, such as HEALPix (Górski et al. 2005), Igloo (Crittenden 2000) or HTM (Szalay et al. 2005), and estimate a piecewise constant prior (uniform in the cells) using the same methodology. Naturally other more sophisticated models can also be used in the same spirit, e.g., specific functional forms or smoothing to limit the gradient, as well as tapered windows when required.

The threshold on the posterior,  $P_T$ , can also be established in a consistent way. Here the requirement is that

$$\sum_{P(H|D) > P_T} 1 = N_* \quad (31)$$

This is equivalent to applying a Bayes classifier. By changing the right hand side of the above equation, it is possible to make the selection more restrictive or less depending on the scientific goal. In the case, where the prior changes on the sky and eq. 30 is solved in cells of some pixelization, one can still just use a single  $P_T$  limit obtained from the entire catalog by ensuring that the total number of objects are consistent. The counts in individual cells may not be perfectly recovered but, if the prior is right, there should be no significant trends.

## 6. PRACTICAL CONSIDERATIONS

The question remains how to evaluate the Bayes factor efficiently for multiple catalogs without considering all possible combinations of sources. Fast algorithms exist to match two sets of point sources using an angular separation limit (Budavári et al. 2003; Malik et al. 2003; Gray et al. 2004, 2006; Szalay et al. 2005; Nieto-Santisteban 2007). Ideally one would like to leverage the power of these two-way cross-match engines in a recursive manner, and get rid of unlikely combinations with small Bayes factors as early as possible.

Matching two catalogs is straightforward; any Bayes factor limit corresponds to a single distance cut, and hence our existing tools are adequate. To go from  $n$  number of catalogs to  $n+1$ , we need to make this process iterative, and prune the match list step-by-step. We do this by computing the overall Bayes factor in every step assuming that all other subsequent catalogs will contribute sources at the best possible position. This optimization problem may be expensive to solve in general, but can be analytically calculated in special cases, and for the spherical normal distribution the solution is evident: the center position of the mode is the correct choice.

In fact, for the normal distribution one can do even better. In every step, a new catalog is added to the current sub-matches. Since the product of normal distributions is still of the same functional form, one can compute the Bayes factor as a function of angular separation from that position, derive the limiting radius, and utilize a two-way crossmatch engine for joining the current  $k$ -tuples with the new  $(k+1)^{\text{th}}$  catalog using that threshold. For this we rewrite the logarithm of the Bayes factor in eq. 18, in the more convenient form of

$$\ln B = \ln N - \frac{1}{2} \sum_{i=2}^n \frac{a_{i-1}}{a_i} w_i \Delta_i^2 \quad (32)$$

with the newly introduced variables

$$N = 2^{n-1} \frac{\prod w_i}{\sum w_i} \quad (33)$$

$$a_k = \sum_{i=1}^k w_i \quad (34)$$

$$\vec{\Delta}_i = \vec{x}_i - \vec{c}_{i-1} \quad (35)$$

where  $\vec{c}_k$  is the unit vector of the best position for the current  $k$ -tuple of sub-match,

$$\vec{c}_k = \sum_{i=1}^k w_i \vec{x}_i / \left| \sum_{i=1}^k w_i \vec{x}_i \right| \quad (36)$$

With these we compute the weight of evidence in a recursive manner. The iteration starts by substituting  $\vec{c}_1 = \vec{x}_1$ . In the  $k^{\text{th}}$  step, the maximum search radius  $\rho_{k+1}$  is computed from eq. 32 to yield the Bayes factor threshold  $B_0$  by assuming optimal matches from the subsequent catalogs with vanishing  $\Delta_i^2$  contributions,

$$b_{k+1}\rho_{k+1}^2 = 2\ln\frac{N}{B_0} - \sum_{i=2}^k b_i\Delta_i^2 \quad \text{with} \quad b_k = \frac{w_k a_{k-1}}{a_k} \quad (37)$$

We assign every source within that radius to each  $k$ -tuple submatch, and go to the next catalog. In general, the search radius will be different for every tuple for their different spatial configurations. When the two-way matching algorithm requires a fixed radius, one can take the maximum value in linear time, use that more generous search radius in the matching, and filter the result set later, just before going to the next catalog. From catalog to catalog we propagate only the quantities that are necessary to calculate the weight of evidence. The recursion formulas are given by the following expressions:

$$a_k = a_{k-1} + w_k \quad (38)$$

$$q_k = q_{k-1} + \frac{a_{k-1}}{a_k} w_k \Delta_k^2 \quad (39)$$

$$\vec{c}_k = \left( \vec{c}_{k-1} + \frac{w_k}{a_k} \vec{\Delta}_k \right) / \left| \vec{c}_{k-1} + \frac{w_k}{a_k} \vec{\Delta}_k \right| \quad (40)$$

This stepwise method for evaluating the weight of evidence not only provides an accurate match list that meets all our requirements enumerated in Section 1, e.g., symmetry in the catalogs, but also exhibits the performance of the current state-of-the-art two-way crossmatching tools.

## 7. SUMMARY

We presented a general probabilistic formalism for cross-identifying astronomical point sources. The framework is

based on Bayesian hypothesis testing to decide whether a series of observations truly belong to a single astronomical object. The expression we derived is completely general, symmetric in all observations, and accommodates any model of the astrometric precision. We introduced the spherical normal distribution, and calculated the Bayes factor for the generic  $n$ -way matching problem both in the general case and in the typical limit of high precision and small angular separations. The cases of 2- and 3-way matching were studied in detail. We discussed an efficient evaluation strategy of the Bayes factor that leverages the power of existing high-performance two-way matching tools in a recursive manner, yet, it provides accurate measurements of the observational evidence that are independent of the order of the catalogs considered. While the normal distribution is the simplest to work with for its unique properties, other specific PDFs can be handled in the same spirit. Our technique provides a natural mechanism to include other observed properties. We demonstrated how multicolor survey data, even at different wavelengths, can be utilized in the matching process by invoking SED models. Morphological classification or redshift measurements, when available, will also increase the accuracy of the results.

The beauty of our approach to the cross-identification problem is that it completely separates the dependence on each parameter, while providing the opportunity to incorporate them in a fairly straightforward way. Including expert knowledge about the physics of the objects in the analysis is easily achievable by adopting the right priors, and when such information is not available, self-consistency arguments can guide the process to a stable solution in a few iterations. With the pre-computed Bayes factors in the matched catalogs, astronomers can define custom thresholds to derive specialized crossmatch catalogs based on their own explicit assumptions. For example, using a database of the same set of associations, researchers can optimize for completeness of the galaxy population, or even search for unusually red objects.

## APPENDIX

### THE BAYES FACTOR AND THE SPHERICAL NORMAL DISTRIBUTION

In this appendix we discuss the mathematical calculation of the Bayes factor in the common case, when a spherical normal distribution is assumed for modelling the astrometric accuracy. In addition we also adopt an all-sky prior in this derivation.

The Bayes factor is the ratio of the likelihoods,  $p(D|H)$  and  $p(D|K)$ , where again  $D$  represents the observed positions,  $\{\vec{x}_i\}$ .

$$B = \frac{p(D|H)}{p(D|K)} \quad (A1)$$

We recall that hypothesis  $H$  is parameterized by a single position,  $\vec{m}$  unit vector, and  $K$  is parameterized by a set of  $n$  position vectors,  $\{\vec{m}_i\}$ . The basic equations to start from are

$$p(D|H) = \int d^3m \, p(\vec{m}|H) \, p(D|\vec{m}, H) \quad (A2)$$

$$p(D|K) = \int d^3m_1 \int d^3m_2 \dots \int d^3m_n \, p(\vec{m}_1|K) p(\vec{m}_2|K) \dots p(\vec{m}_n|K) \, p(D|\{\vec{m}_i\}, K) \quad (A3)$$

where

$$p(\vec{m}|M) = \frac{\delta(|\vec{m}|-1)}{4\pi} \quad (A4)$$

$$p(\{\vec{x}_i\}|\vec{m}, H) = \prod_i^n N(\vec{x}_i|\vec{m}, w_i) = \prod_i^n \frac{w_i \delta(|\vec{x}_i|-1)}{4\pi \sinh w_i} \exp(w_i \vec{x}_i \vec{m}) \quad (A5)$$

$$p(\{\vec{x}_i\}|\{\vec{m}_i\}, K) = \prod_i^n N(\vec{x}_i|\vec{m}_i, w_i) = \prod_i^n \frac{w_i \delta(|\vec{x}_i|-1)}{4\pi \sinh w_i} \exp(w_i \vec{x}_i \vec{m}_i) \quad (A6)$$

First we focus on hypothesis  $H$

$$p(D|H) = \int d^3m \frac{\delta(|\vec{m}|-1)}{4\pi} \prod_i^n \frac{w_i \delta(|\vec{x}_i|-1)}{4\pi \sinh w_i} \exp(w_i \vec{x}_i \vec{m}) \quad (\text{A7})$$

$$= \left( \prod_i^n \frac{w_i \delta(|\vec{x}_i|-1)}{4\pi \sinh w_i} \right) \int d^3m \frac{\delta(|\vec{m}|-1)}{4\pi} \exp\left(\sum_i^n w_i \vec{x}_i \vec{m}\right) \quad (\text{A8})$$

introduce

$$w\vec{x} = \sum_i^n w_i \vec{x}_i \quad (\text{A9})$$

where  $\vec{x}$  is a unit vector, and write

$$p(D|H) = \left( \prod_i^n \frac{w_i \delta(|\vec{x}_i|-1)}{4\pi \sinh w_i} \right) \int d^3m \frac{\delta(|\vec{m}|-1)}{4\pi} \exp(w\vec{x}\vec{m}) \quad (\text{A10})$$

$$= \left( \frac{\sinh w}{w} \prod_i^n \frac{w_i \delta(|\vec{x}_i|-1)}{4\pi \sinh w_i} \right) \int d^3m \frac{w \delta(|\vec{m}|-1)}{4\pi \sinh w} \exp(w\vec{x}\vec{m}) \quad (\text{A11})$$

$$= \frac{\sinh w}{w} \prod_i^n \frac{w_i}{\sinh w_i} \frac{\delta(|\vec{x}_i|-1)}{4\pi} \quad (\text{A12})$$

The likelihood of the alternative hypothesis  $K$  is calculated similarly

$$p(D|K) = \prod_i^n \int d^3m_i \frac{\delta(|\vec{m}_i|-1)}{4\pi} \frac{w_i \delta(|\vec{x}_i|-1)}{4\pi \sinh w_i} \exp(w_i \vec{x}_i \vec{m}_i) \quad (\text{A13})$$

$$= \prod_i^n \frac{\delta(|\vec{x}_i|-1)}{4\pi} \quad (\text{A14})$$

Hence the Bayes factor is

$$B = \frac{\sinh w}{w} \prod_i^n \frac{w_i}{\sinh w_i} \quad (\text{A15})$$

as also shown in eq. 13.

## HIGH ASTROMETRIC ACCURACY AND SMALL SEPARATIONS

The astrometric precision of the actual observations is almost always extremely high in the absolute sense, so it is worth examining the approximation of the Bayes factor in this limit. We also assume small angular separations. In the chain of equations below we only use the “ $\approx$ ” sign to signal new approximations. We start from the previous result

$$B = \frac{\sinh w}{w} \prod_i^n \frac{w_i}{\sinh w_i} \quad (\text{B1})$$

$$\approx 2^{n-1} \frac{e^w}{w} \prod_i^n \frac{w_i}{e^{w_i}} \quad (\text{B2})$$

$$= 2^{n-1} \frac{\prod_i^n w_i}{w} e^{w - \sum w_i} \quad (\text{B3})$$

$$= 2^{n-1} \frac{\prod_i^n w_i}{w} e^{\sum w_i \left(\frac{w}{\sum w_i} - 1\right)} \quad (\text{B4})$$

where we exploit the fact that all  $w$  weights are large, hence the  $\sinh w$  is approximately  $\frac{1}{2} \exp w$ . We proceed by calculating

$$\left( \frac{w}{\sum_i w_i} \right)^2 = \frac{w^2}{(\sum_i w_i)^2} \quad (\text{B5})$$

$$= \frac{\sum_i w_i^2 + 2 \sum_{i < j} w_i w_j \vec{x}_i \vec{x}_j}{\sum_i w_i^2 + 2 \sum_{i < j} w_i w_j} \quad (\text{B6})$$

$$= \frac{\sum_i w_i^2 + 2 \sum_{i < j} w_i w_j \cos \psi_{ij}}{\sum_i w_i^2 + 2 \sum_{i < j} w_i w_j} \quad (\text{B7})$$

$$\approx \frac{\sum_i w_i^2 + 2 \sum_{i < j} w_i w_j (1 - \psi_{ij}^2/2)}{\sum_i w_i^2 + 2 \sum_{i < j} w_i w_j} \quad (\text{B8})$$

$$= \frac{\sum_i w_i^2 + 2 \sum_{i < j} w_i w_j - \sum_{i < j} w_i w_j \psi_{ij}^2}{\sum_i w_i^2 + 2 \sum_{i < j} w_i w_j} \quad (\text{B9})$$

$$= 1 - \frac{\sum_{i < j} w_i w_j \psi_{ij}^2}{(\sum_i w_i)^2} \quad (\text{B10})$$

After taking the square root of the above equation, we get

$$\frac{w}{\sum_i w_i} \approx 1 - \frac{\sum_{i < j} w_i w_j \psi_{ij}^2}{2 (\sum_i w_i)^2} \quad (\text{B11})$$

and

$$\sum_i w_i \left( \frac{w}{\sum_i w_i} - 1 \right) = - \frac{\sum_{i < j} w_i w_j \psi_{ij}^2}{2 \sum_i w_i} \quad (\text{B12})$$

From the above equations we also see that

$$\frac{1}{w} \approx \frac{1}{\sum_i w_i} \left( 1 + \frac{\sum_{i < j} w_i w_j \psi_{ij}^2}{2 (\sum_i w_i)^2} \right) \approx \frac{1}{\sum_i w_i} \quad (\text{B13})$$

in this context to only keep the leading term. By substituting the above two equations to eq. B4, we arrive at our generic small angle result shown in eq. 18. The 2- and 3-way matching cases are straightforward specializations of the generic equation, where one substitutes  $w_i = 1/\sigma_i^2$  to work out the simplified formulae.

The authors are grateful for invaluable discussions with María Nieto-Santisteban, István Csabai and Zoltán Rác on various aspects of the topic, and gladly acknowledge the generous support from the following organizations: Gordon and Betty Moore Foundation GBMF 554, W. M. Keck Foundation KECK D322197, NSF NVO AST-0122449, NASA AISRP NNG05GB01G, NASA GALEX 44G1071483, Hungarian National Scientific Foundation OTKA-T047244, European Research Training Network MRTN-CT-2004-503929. Part of this research was done while A. S. was a recipient of an Alexander von Humboldt Fellowship at the MPA.

#### REFERENCES

- Breitenberger, E. 1963, *Biometrika*, 50, 81  
 Bruzual, G., & Charlot, S. 2003, *MNRAS*, 344, 1000  
 Budavári, T., Malik, T., Szalay, A. S., Thakar, A. R., & Gray, J. 2003, in *ASP Conf. Ser.*, Vol. 295, *Astronomical Data Analysis Software and Systems (ADASS) XII*, eds. H. E. Payne, R. I. Jedrzejewski, and R. N. Hook (San Francisco: ASP), p.31  
 Budavári, T., et al. 2004, in *ASP Conf. Ser.*, Vol. 314, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, eds. F. Ochsenbein, M. Allen, and D. Egret (San Francisco: ASP), p.177  
 Crittenden, R. G. 2000, *Astrophysical Letters Communications*, 37, 377  
 Fisher, R., 1953, *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, Vol. 217, No. 1130., pp.295–305  
 Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., & Bartelmann, M. 2005, *ApJ*, 622, 759  
 Gray, J., Szalay, A. S., Fekete, G., Nieto-Santisteban, M. A., O'Sullivan, W., Thakar, A. R., Heber, G., & Rots, A. H. 2004, Microsoft Research Technical Report, MSR-TR-2004-32  
 Gray, J., Nieto-Santisteban, M. A., & Szalay, A. S. 2006, Microsoft Research Technical Report, MSR-TR-2006-52  
 Lévy, P. 1939, *Bull. Soc. Math. Fr.*, 67, 1  
 Madore, B. F., Helou, G., Corwin, H. G., Jr., Schmitz, M., Wu, X., & Bennett, J. 1992, *Astronomical Data Analysis Software and Systems I*, 25, 47  
 Malik, T., Szalay, A. S., Budavári, T., & Thakar, A. R. 2003, *Proceedings of Conference on Innovative Data Systems Research (CIDR)*, 17  
 Mann, R. G., et al. 1997, *MNRAS*, 289, 482  
 Martin, D. C., et al. 2005, *ApJ*, 619, L1  
 Morrissey, P., et al. 2007, *ApJS* in press, *ArXiv e-prints*, 706 (arXiv:0706.0755)  
 Nieto-Santisteban, M. A. 2007, PhD thesis, in preparation  
 Ochsenbein, F., Bauer, P., & Marcout, J. 2000, *A&AS*, 143, 23  
 Pier, J. R., et al. 2003, *AJ*, 125, 1559  
 Rohde, D. J., Gallagher, M. R., Drinkwater, M. J., & Pimblett, K. A. 2006, *MNRAS*, 369, 2  
 Sutherland, W., & Saunders, W. 1992, *MNRAS*, 259, 413  
 Szalay, A. S., Gray, J., Fekete, G., Kunszt, P., Kukol, P., & Thakar, A. R. 2005, Microsoft Research Technical Report, MSR-TR-2005-123  
 York, D. G., et al. 2000, *AJ*, 120, 1579